A modified two-stage method for parameter estimation in sinusoidal models of correlated gene expression profiles

Thailand Statistician
January 2020; 18(1): 77-89
http://statassoc.or.th
Contributed paper

A Modified Two-Stage Method for Parameter Estimation in Sinusoidal Models of Correlated Gene Expression Profiles

Wannapa Pukdee [a], Orathai Polsen*[a] and Mohamed Fazil Baksh [b]

[a] Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand.

[b] Department of Mathematics and Statistics, School of Mathematical, Physical and Computational Sciences, University of Reading, Reading, United Kingdom.

*Corresponding author; e-mail: orathai.p@sci.kmutnb.ac.th

Received: 13 June 2018
Revised: 1 November 2018
Accepted: 13 May 2019

Abstract

A two-stage method by Seber and Wild (2003) used to fit nonlinear regression models with correlated errors by using residuals obtained from the ordinary least square estimation has been shown by Pukdee et al. (2018) to underestimate the standard errors of parameter estimates in sinusoidal models, leading to poor coverage probabilities. In order to improve inferential statistics, a modified two-stage method is developed using residuals from the one-way ANOVA model to estimate variance components in the iterative estimation procedure and compared with the two-stage, conditional least squares and generalized least squares methods. A simulation study shows that the proposed method has similar successful convergence rates as the two-stage and conditional least squares methods but produces more reliable point and interval estimates. Although very little difference is seen between estimates produced by generalized least squares and the proposed method, the latter has a consistently higher successful convergence rate, and consequently is more likely to produce a result than the former, and this difference in rates becomes substantial when the model complexity increases.

Keywords: Non-linear regression, correlated responses, two-stage method, generalized least squares.

1. Introduction

Sinusoidal functions are used in modelling data displaying a cyclic pattern over time, such as obtained in studies on circadian rhythms of biological organisms, where reliable estimates of model parameters, such as the frequency, are required. Circadian rhythms are regulators of many biological processes and are studied within pharmaceutics as they can be useful predictors of drug metabolism, dosage and efficacy. Gene expression, the process by which information from a gene is used in the synthesis of a functional product, is measured using bioluminescence technology. The responses arising from the study of circadian gene expression are measurements of light intensity over time. Typically data is collected on the same experimental unit at selected time points over a period of time.

While de-trending (Kyriacou and Hall 1980; Izumo et al. 2003; Izumo et al. 2006; Maier et al. 2009; Yang and Su 2010) is a widely used technique to fit sinusoidal correlated data models to correlated gene expression data, recent work (Pukdee et al. 2018) has shown that de-trending leads to biased parameter estimates compared to conditional least squares (Bates and Watts 1988) and a two-stage estimation approach (Seber and Wild 2003). However, Pukdee et al. (2018) has also shown that both the two-stage (TS) and conditional least squares (CLS) methods tend to underestimate the standard errors of parameter estimators as model complexity increases and when the correlation between adjacent responses is high. An alternative to TS and CLS is generalized least squares (GLS) estimation (Davidian and Giltinan 1995). The above three estimation methods utilize the least squares procedure and so can potentially benefit from the standard distributional properties of least squares estimators but GLS is well known to face convergence problems when fitting complicated regression models of correlated data. In this paper, the issues of a more accurate variance estimator and successful convergence of the nonlinear iterative procedure is addressed by proposing a modified two-stage (MTS) estimation method that uses the residuals from the one-way ANOVA model of replicate observations at each time point. The proposed method is developed and compared to GLS, CLS and TS methods in this paper.

2. Methods

The nonlinear regression model of the relationship between an independent variable t, here time, and a dependent response variable y measured at n time points for each of r experimental units is

$$\mathbf{y}_i = \mathbf{f}(\mathbf{t}_i;\boldsymbol{\theta}) + \boldsymbol{\varepsilon}_i;\ i = 1,\ldots,r, \tag{1}$$

where $\mathbf{y}_i = (y_{i,1},\ldots,y_{i,n})'$ denotes the observed response vector of the i^{th} unit, $\mathbf{t}_i = (t_{i,1},...,t_{i,n})'$ is the vector of time points, $\mathbf{f}(\mathbf{t}_i;\boldsymbol{\theta}) = (f(t_{i,1};\boldsymbol{\theta}),\ldots,f(t_{i,n};\boldsymbol{\theta}))'$ is some nonlinear function f of t and an unknown parameter vector $\boldsymbol{\theta}$, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1},\ldots,\varepsilon_{i,n})'$ is a vector of correlated errors. Assuming the repeated measures on each experimental unit follows a stationary autoregressive process of order 1, AR(1), the error components can be described as a linear relationship between terms at time points j and $j-1$ by

$$\varepsilon_{i,j} = \rho\varepsilon_{i,j-1} + \delta_{i,j};\ j = 1,\ldots,n, \tag{2}$$

where $\rho \in [-1,1]$ is the correlation coefficient for $\varepsilon_{i,j}$ and $\varepsilon_{i,j-1}$, and $\delta_{i,j}$ are independent and identically distributed (i.i.d.) variables with zero mean and constant variance σ_δ^2. Under this model, $\varepsilon_{i,j}$ have mean 0 and variance $\sigma^2 = \dfrac{\sigma_\delta^2}{1-\rho^2}$.

In this paper, four sinusoidal nonlinear functions found in the literature (Kyriacou and Hall 1980; Izumo et al. 2003; Izumo et al. 2006; Maier et al. 2009) and relevant for modelling circadian data are evaluated. The first is the one-sine function

$$f(t;\boldsymbol{\theta}) = \alpha + \beta t + a\exp(-dt)\sin\left(\frac{2\pi t}{\tau} + \Phi\right),$$

where τ is the period, a is the amplitude, Φ represents the phase of the sine wave, d is a damping parameter, α is an intercept and β is a slope of the linear trend. Secondly, the song-sine function is

$$f(t;\boldsymbol{\theta}) = \alpha + \beta t + (a_s + a\exp(-dt))\sin\left(\frac{2\pi t}{\tau} + \Phi\right),$$

where a_s is a linear constant displacement in the amplitude. The third is the two-sine with damping function to deal with the potential of more than one sinusoidal pattern,

$$f(t;\boldsymbol{\theta}) = \alpha + \beta t + a\exp(-dt)\sin\left(\frac{2\pi t}{\tau} + \Phi\right) + b\sin\left(\frac{2\pi t}{\upsilon} + \Phi\right),$$

where b and υ are the amplitude and the period respectively of the second sine term, and is proposed as a novel function. Fourthly is the two-sine without damping function, also used to describe circadian patterns with two different periods,

$$f(t;\boldsymbol{\theta}) = \alpha + \beta t + a\sin\left(\frac{2\pi t}{\tau} + \Phi\right) + b\sin\left(\frac{2\pi t}{\upsilon} + \Phi\right).$$

The two-sine function comprises two amplitudes which are assumed to be significantly different from zero and are extensions of the one-sine function provided above.

The above nonlinear regression models with correlated errors are fitted in this paper using conditional least squares, a two- stage estimation approach, generalized least squares and a new modified two- stage approach. As explained below, all four methods are taking account of the correlation structure in the data in different ways.

2.1. The generalized least squares method

In the situation where the error term $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1},\ldots,\varepsilon_{i,n})'$ for subject i is serially correlated and assumed to be a stationary AR(1) process, $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0},\sigma^2\boldsymbol{V}_i)$ where

$$\boldsymbol{V}_i = \begin{bmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{n-1} \\ \rho & 1 & \rho & \ldots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \ldots & \rho^{n-3} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \ldots & 1 \end{bmatrix}.$$

The GLS estimator is obtained by minimizing the error sum of squares

$$\left\{\mathbf{y}_i - \mathbf{f}(\mathbf{t}_i;\boldsymbol{\theta})\right\}' \boldsymbol{V}_i^{-1} \left\{\mathbf{y}_i - \mathbf{f}(\mathbf{t}_i;\boldsymbol{\theta})\right\}.$$

In cases where the GLS method fails to converge when using iteratively reweighted least squares for parameter estimation (Seber and Wild 2003), a transformation can be considered. Since $\boldsymbol{V}_i$ is a positive definite matrix, then there exists an upper triangular matrix $\boldsymbol{U}_i$ such that $\boldsymbol{V}_i = \boldsymbol{U}_i'\boldsymbol{U}_i$ and $\boldsymbol{V}_i^{-1} = \boldsymbol{R}_i'\boldsymbol{R}_i$, as defined by

$$\boldsymbol{R}_i = \begin{bmatrix} \sqrt{1-\rho^2} & 0 & 0 & \ldots & 0 & 0 \\ -\rho & 1 & 0 & \ldots & 0 & 0 \\ 0 & -\rho & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -\rho & 1 \end{bmatrix},$$

where $\boldsymbol{R}_i = \left(\boldsymbol{U}_i'\right)^{-1}$. Note that Cholesky factorization aims to calculate the matrix $\boldsymbol{U}_i$. Applying the Cholesky decomposition transforms the model to an ordinary nonlinear least squares model. The GLS method is implemented by using iterative maximum likelihood estimation for the mean $\boldsymbol{\theta}$ and

variance components in $\boldsymbol{V}_i$ (Pinheiro and Bates 2000). An empirical autocorrelation function is used as starting value for ρ in the iterative procedure.

2.2. The two-stage method

The two-stage (TS) approach, proposed by Gallant and Goebel (1976), is a version of the GLS method for estimating the variance components in $\boldsymbol{V}_i$. The TS method under the same assumption as the above GLS approach for estimating parameters of a nonlinear time-series regression with AR(1) errors, consists of two ordinary least squares (OLS) procedures. In the first stage, the correlation structure is ignored and the model (1) is fitted by OLS to produce estimates $\hat{\boldsymbol{\theta}}_{\text{OLS}}$ and fitted values $f(t_{i,j};\hat{\boldsymbol{\theta}}_{\text{OLS}})$. The residual vector for the i^{th} unit

$$\hat{\boldsymbol{\varepsilon}}_i = \mathbf{y}_i - \mathrm{f}(\mathbf{t}_i;\hat{\boldsymbol{\theta}}_{\text{OLS}}),$$

is then calculated and used to produce an estimate of the within subject correlation ρ_i (Park and Mitchell 1980) given by

$$\hat{\rho}_i = \frac{\sum_{j=2}^{n} \hat{\varepsilon}_{i,j}\hat{\varepsilon}_{i,j-1}}{\sum_{j=2}^{n-1} \hat{\varepsilon}_{i,j}^2}. \tag{3}$$

In the second stage, using the mean of the *r* estimates obtained $\hat{\rho}_1,\ldots,\hat{\rho}_r$, denoted $\hat{\rho}$, to estimate the (assumed) common correlation ρ, a transformed model is expressed in matrix form with i.i.d. errors $\boldsymbol{\delta}_i \sim (\mathbf{0},\sigma_\delta^2\boldsymbol{I}_i)$ as

$$\mathbf{z}_i = \mathbf{g}(\mathbf{t}_i;\boldsymbol{\theta}) + \boldsymbol{\delta}_i;\ i = 1,\ldots,r,$$

where $\mathbf{z}_i = \hat{\boldsymbol{R}}_i\mathbf{y}_i$, $\mathbf{g}(\mathbf{t}_i;\boldsymbol{\theta}) = \hat{\boldsymbol{R}}_i\mathrm{f}(\mathbf{t}_i;\boldsymbol{\theta})$, $\boldsymbol{\delta}_i = \hat{\boldsymbol{R}}_i\boldsymbol{\varepsilon}_i$, where $\hat{\boldsymbol{R}}_i$ is the estimate of $\boldsymbol{R}_i$ formed by replacing ρ with $\hat{\rho}$. As the matrix $\hat{\boldsymbol{R}}_i$ is constructed and fitted using OLS, the TS procedure is very simple to code and implement. Gallant and Goebel (1976) improved the estimator of the two-stage method by repeating the above procedure. In this repeat, the residuals $\hat{\boldsymbol{\varepsilon}}_i = \mathbf{y}_i - \mathrm{f}(\mathbf{t}_i;\hat{\boldsymbol{\theta}}_{\text{TS}})$, where $\hat{\boldsymbol{\theta}}_{\text{TS}}$ is the two-stage estimator obtained in the first implementation, are used to obtain a new estimate of the correlation in the weight matrix. Additionally, the TS procedure produces estimators with asymptotic properties similar to OLS estimators (Gallant and Goebel 1976).

2.3. The modified two-stage method

Asikgil and Erar (2009) estimated the correlation coefficient in the above weight matrix $\boldsymbol{R}_i$ by using different procedures. Following this idea, the first step in the modified two-stage method estimates the errors ε_{ij} by again ignoring the correlation structure but now fitting a one-way ANOVA model of the replicate observations at each time point. The one-way ANOVA model with i.i.d. errors fitted is

$$y_{ij} = \mu_j + \varepsilon_{ij}, \quad i = 1,\ldots,r;\ j = 2,\ldots,n,$$

where μ_j is the mean response at the j^{th} time point (group).

The residuals,

$$\hat{\varepsilon}_{ij} = y_{ij} - \hat{y}_{j,\text{iid}},$$

where $\hat{y}_{j,\text{iid}}$ is the sample mean for the j^{th} time point, is next used to estimate the correlation coefficient for the i^{th} experimental unit $\hat{\rho}_i$ used in (3). The second stage of the analysis proceeds by using $\hat{\rho} = \dfrac{\hat{\rho}_1 + \ldots + \hat{\rho}_r}{r}$ in the matrix $\boldsymbol{R}_i$ in the above two-stage process. This pure error estimate of ε_{ij} is model independent and therefore likely to be an improvement over any model dependent estimate.

2.4. The conditional least squares method

The conditional least squares (CLS) model is constructed by assuming that the correlated errors $\varepsilon_{i,j}$ are a stationary AR(1) process, as provided by (2), and subtracting ρ times the model for $y_{i,j-1}$ from the model for $y_{i,j}$. This is given by Bates and Watts (1988) as

$$y_{i,j} - \rho y_{i,j-1} = f(t_{i,j};\boldsymbol{\theta}) - \rho f(t_{i,j-1};\boldsymbol{\theta}) + \delta_{i,j};\ \ j = 2,\ldots,n, \tag{4}$$

where $\boldsymbol{\theta}$ is a parameter vector and ρ is the parameter of the AR(1) model, which are estimated by least squares. The CLS approach is implemented by minimizing

$$S(\boldsymbol{\theta},\rho) = \sum_{i=1}^{r}\sum_{j=2}^{n}\left(y_{i,j} - \rho y_{i,j-1} - f(t_{i,j};\boldsymbol{\theta}) + \rho f(t_{i,j-1};\boldsymbol{\theta})\right)^2,$$

with respect to $\boldsymbol{\theta}$ and ρ, jointly. A benefit of this approach is that the estimates obtained are consistent and asymptotically normal (Klimko and Nelson 1978). In addition, Pukdee et al. (2018) shows that CLS produces less biased estimates and more reliable confidence intervals than the TS method when used to analyze circadian rhythms in gene expression profiles. However, the CLS method can increase the risk of lack of convergence in the iterative fitting process due to the fact that the number of parameters in the model (4) increases and the degrees of freedom is reduced by the first order of the autoregressive process. A starting value for ρ in the CLS iterative routine can be obtained by fitting the nonlinear model assuming uncorrelated errors and calculating the residual autocorrelation function (Bates and Watts 1988). Note that it is very important that the starting values should be close to the final parameter estimates to increase the chance of convergence.

3. Simulation Study

To evaluate the performance of the above methods datasets are first generated for different levels of the correlation ρ in an AR(1) process under the conditional least squares model,

$$y_{i,j} = \begin{cases} f(t_{i,j};\boldsymbol{\theta}) + \delta_{i,j} & ;\ j = 1 \\ \rho y_{i,j-1} + f(t_{i,j};\boldsymbol{\theta}) - \rho f(t_{i,j-1};\boldsymbol{\theta}) + \delta_{i,j} & ;\ j = 2,\ldots,n, \end{cases}$$

where $\delta_{i,j}$ are independent and identically distributed $N(0,\sigma^2)$ and $f(t_{i,j};\boldsymbol{\theta})$ is a sinusoidal nonlinear function. For each level of the correlation ρ (0, 0.1, 0.25, 0.5, 0.75, 0.9) with $\sigma_\delta^2 = 25$ a total of 10,000 replicate studies each are generated under the one-sine, song-sine and two-sine models with parameter values $\boldsymbol{\theta}$ as provided in Table 1. For each simulation study, repeated measures are simulated for $r = 4$ independent subjects at times $t_{i,j} = 0, 1.5, \ldots, 78$ and $n = 53$.

Table 1 The four sets of parameter values used in the simulations

Model	θ								
	τ	υ	a_s	a	b	Φ	d	α	β
one-sine	24	-	-	180	-	0.31	0.07	330	-3
song-sine	24	-	0.5	180	-	0.31	0.07	330	-3
two-sine with damping	24	35	-	180	0.5	0.31	0.07	330	-3
two-sine without damping	24	35	-	180	0.5	0.31	-	330	-3

Each simulated dataset is analysed by fitting the sinusoidal regression models in Table 1 using the GLS, TS, MTS and CLS methods described above. The R software (R Core Team 2013) with the nls function and the nlme library, see Pinheiro and Bates (2000), Ritz and Streibig (2008), and Crawley (2013), is used to fit the models. Estimates of bias, mean square error and coverage probability are next obtained and used to compare accuracy and efficiency of the period estimator, as well as accuracy of the period variance estimator for the four methods. The percentage bias of the estimator is

$$\%\text{Bias} = 100\left(\frac{\text{Bias}(\hat{\tau})}{\tau}\right),$$

where $\text{Bias}(\hat{\tau}) = \hat{\tau} - \tau$; $\hat{\tau}$ is the mean of $\hat{\tau}_m$ and $\hat{\tau}_m$ is the period estimate obtained from the m^{th} simulation run $(m = 1, 2\ldots, M)$. In order to assess the precision of the estimated standard error for parameter estimates, the percentage relative difference between the standard deviation and the standard error for the estimate is given by

$$\%\text{Diff} = 100\left(\frac{\text{SE}(\hat{\tau}) - \text{SD}(\hat{\tau})}{\text{SD}(\hat{\tau})}\right),$$

where $\text{SD}(\hat{\tau}) = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(\hat{\tau}_m - \hat{\tau})^2}$ and $\text{SE}(\hat{\tau}) = \frac{1}{M}\sum_{m=1}^{M}\text{SE}(\hat{\tau}_m)$, with $\text{SE}(\hat{\tau}_m)$ the estimated standard error of the period estimate from the m^{th} simulated dataset while the root mean square error is estimated by

$$\text{RMSE} = \sqrt{\left(\text{SD}(\hat{\tau})\right)^2 + \left(\text{Bias}(\hat{\tau})\right)^2}.$$

The estimated coverage probability is provided by the proportion of times that the $100(1-\alpha)\%$ confidence interval (CI) covers the true value of τ, which is given by

$$\hat{\tau}_m \pm t_{\frac{\alpha}{2},v}\,\text{SE}(\hat{\tau}_m),$$

where $t_{\frac{\alpha}{2},v}$ is the critical value of student t distribution with the significance level α and v degrees of freedom.

Convergence of the iterative algorithms was not achieved in all instances for all the methods and for all the fitted models. Many failures would lead to less precise simulation results (Burton et al. 2006) and mitigates against utility of the method in practice. Provided in Table 2 is the achieved percentage of successful convergences from 10,000 replications for each method when the one-sine, song-sine, two-sine with and without damping models are fitted.

It can clearly be seen that the convergence rates of the four methods vary considerably and seem to decrease with increasing model complexity. Under the simplest one-sine model, simulation results with TS and MTS methods show 100% successful fits for all ρ and CLS and also gives full

successful convergences for moderate and high ρ. Under the two-sine with damping model, the three methods, CLS, TS and MTS have similar failure to convergence rates of less than 8%, while the GLS method failed in approximately 10%-18% of the time. For the two-sine without damping model, MTS has the best successful convergence rate, while GLS failure to converge rate for moderate $(\rho = 0.5)$ and high $(\rho = 0.75, 0.90)$ correlations is around 24%-30%. The lowest convergence rates of around 45%-65% are observed for the song-sine model fitted by GLS.

Provided in Table 3 are estimates of the bias in estimation of the period τ (%Bias), bias in estimation of the standard error for $\hat{\tau}$ (%Diff) and root mean square error (RMSE) when the CLS, TS, GLS and MTS methods are used to fit the one-sine, song-sine, two-sine with and without damping models; provided in Figure 1 are estimates of the corresponding 95% coverage probabilities. It can be seen from Table 3 that while %Bias for all four methods are comparable under all scenarios, the same cannot be said for the bias in estimating the standard error (%Diff). When the one-sine model is fitted, the MTS and GLS methods provide comparatively unbiased variance estimates relative to the CLS and TS methods. Notwithstanding the poor convergence rate for GLS seen earlier in Table 2, this is also largely true when the song-sine model is fitted and is reflected in the coverage probabilities from the MTS and GLS methods being close to the nominal rate of 95%, as depicted in Figure 1(a) and Figure 1(b). When the fitted model is two-sine with damping, %Diff of MTS and GLS are again similar and also their coverage probability are close to 95% at $\rho = 0.00$ and $\rho = 0.10$, but GLS produces slightly better coverage probabilities than those MTS for $\rho = 0.25,\ 0.50, 0.75$ and 0.90, as shown in Figure 1(c), albeit with a much higher failure to converge rate of approximately 15% at $\rho = 0.75$ and 18% at $\rho = 0.90$ (see Table 2). When the two-sine without damping model is fitted, the standard error of $\hat{\tau}$ are underestimated by CLS, TS and MTS methods when $\rho = 0.75$ with values of %Diff around -63%, -47%, -42%, respectively, while the one by GLS method is overestimated with %Diff of approximately 11. In addition, coverage probabilities of GLS are over 95% for high ρ, but GLS and MTS produce coverage probabilities close to 95% for small and moderate ρ in Figure 1(d). Moreover, in terms of RMSE, the three efficient methods are GLS, MTS and TS for fitting all three models, one-sine, song-sine and two-sine with damping, while CLS is considerably less efficient. For fitting the last two-sine without damping, GLS is the best choice, but TS and MTS can be comparable for $\rho = 0.25$ and 0.75.

Table 2 Achieved percentage of successful convergences of the CLS, TS, GLS and MTS iterative methods when fitting sinusoidal regression models

Fitted model	ρ	CLS	TS	GLS	MTS
one-sine	0.00	98.97	100.00	99.62	100.00
	0.10	99.46	100.00	99.33	100.00
	0.25	100.00	100.00	98.45	100.00
	0.50	100.00	100.00	91.96	100.00
	0.75	100.00	100.00	75.64	100.00
	0.90	100.00	100.00	74.58	100.00
song-sine	0.00	98.63	97.93	44.59	99.06
	0.10	98.21	98.93	45.76	98.98
	0.25	98.51	98.70	46.96	99.02
	0.50	98.36	98.78	51.98	99.09
	0.75	98.85	97.76	61.31	98.89
	0.90	97.64	98.85	66.14	98.77
two-sine with damping	0.00	97.21	97.06	91.40	98.11
	0.10	97.22	96.72	91.56	97.67
	0.25	97.55	95.73	90.25	97.44
	0.50	97.20	94.35	88.07	96.80
	0.75	96.67	93.11	85.66	96.39
	0.90	96.21	92.30	82.87	95.90
two-sine without damping	0.00	90.92	90.67	87.69	92.73
	0.10	90.18	89.42	86.02	91.39
	0.25	88.90	85.88	83.22	89.14
	0.50	85.00	79.64	76.86	86.07
	0.75	83.61	74.04	70.27	83.55
	0.90	82.53	71.77	67.71	82.86

4. Example Study

The methods described and evaluated in the previous section can be applied to many research studies in the biological, chemical and physical sciences. An example provided here is a study of a preclinical investigation in drug development in a pharmaceutical company. The responses arise from the study of circadian gene expression as part of the results of an experiment run over 78 hours. The same treatment was applied to four different sets of cells. Each cell is measured every 1.5 hours. The responses oscillate in a similar manner. The repeated responses are measured on the condition that no effects at 0 h are removed. The observations on different cells are assumed independent.

The four models described in Section (2), with an AR (1) covariance structure were fitted to the data using the methods described. As mentioned before, nonlinear regression estimation is based on an iterative algorithm with initial values setting for $\boldsymbol{\theta}$ in Table 1 and ρ by using the mean of $\hat{\rho}_i$ in (3) in which the residuals come from the nonlinear model fitted by OLS, but for MTS the residuals are obtained from fitting the one-way ANOVA model. Table 4 summaries the analyses in terms of the 95% confidence interval (CI) for τ and the standard error estimate, $\hat{\sigma} = \sqrt{\dfrac{\sum_{i=1}^{r}\sum_{j=1}^{n}\left(y_{ij} - \mathrm{f}(t_{ij};\hat{\boldsymbol{\theta}})\right)^2}{\nu}}$, obtained using CLS, TS, GLS and MTS approaches.

Table 3 Estimates of bias of the period parameter τ (%Bias), bias of the standard error for $\hat{\tau}$ (%Diff) and root mean square error (RMSE) obtained when the CLS, TS, GLS and MTS iterative methods are used for fitting sinusoidal regression models

Fitted model	ρ	CLS	TS	GLS	MTS
		%Bias			
one-sine	0.00	0.0069	0.0070	0.0073	0.0070
	0.25	-0.0017	0.0003	0.0007	0.0002
	0.75	0.0245	0.0213	0.0163	0.0175
song-sine	0.00	0.0052	0.0063	-0.0061	0.0064
	0.25	-0.0042	-0.0006	-0.0261	-0.0005
	0.75	0.0318	0.0238	-0.0300	0.0219
two-sine with damping	0.00	-0.0189	-0.0371	-0.0444	-0.0350
	0.25	-0.0023	-0.0432	-0.0509	-0.0405
	0.75	0.1182	0.0288	0.0182	0.0275
two-sine without damping	0.00	0.0020	0.0019	0.0023	0.0014
	0.25	0.0049	0.0046	0.0053	0.0040
	0.75	0.0076	0.0023	0.0056	0.0045
		%Diff			
one-sine	0.00	-1.8268	-3.0557	0.1191	0.1104
	0.25	-2.5229	-5.4038	-0.5034	-1.0383
	0.75	-4.0626	-7.0672	0.3510	-0.5537
song-sine	0.00	-2.7240	-3.7396	-2.7810	-0.2430
	0.25	-3.7935	-6.2894	-3.8361	-1.5411
	0.75	-8.2721	-9.0344	-3.2148	-2.0350
two-sine with damping	0.00	-3.4097	-5.0050	-0.6981	-0.7687
	0.25	-5.0557	-8.5044	-2.5417	-3.0373
	0.75	-8.6243	-13.3056	-5.2215	-6.2059
two-sine without damping	0.00	5.0997	2.3367	4.0230	3.3825
	0.25	-7.7560	-1.2290	4.9365	-12.2145
	0.75	-63.7119	-47.8749	11.1269	-42.6787
		RMSE			
one-sine	0.00	0.1360	0.1111	0.1109	0.1111
	0.25	0.1864	0.1371	0.1368	0.1370
	0.75	0.3994	0.2274	0.2308	0.2276
song-sine	0.00	0.1340	0.1096	0.1117	0.1095
	0.25	0.1836	0.1351	0.1382	0.1351
	0.75	0.4022	0.2259	0.2315	0.2259
two-sine with damping	0.00	0.1427	0.1197	0.1187	0.1193
	0.25	0.1980	0.1503	0.1491	0.1502
	0.75	0.4594	0.2643	0.2644	0.2655
two-sine without damping	0.00	0.0204	0.0144	0.0136	0.0147
	0.25	0.0226	0.0183	0.0168	0.0187
	0.75	0.0947	0.0687	0.0268	0.0728

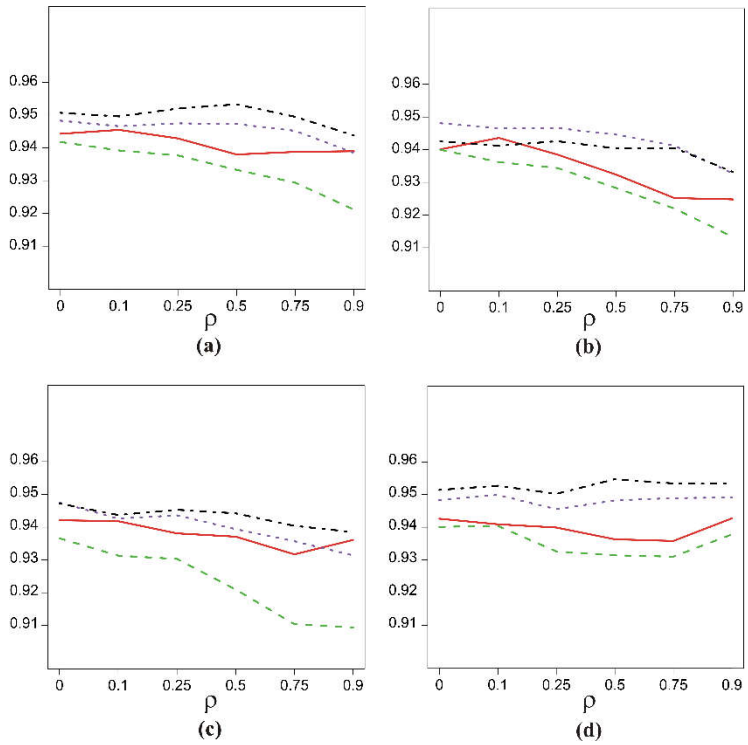

Figure 1 Plots of coverage probability of 95% confidence interval for the period τ using CLS (solid line), TS (dashed line), MTS (dotted line) and GLS (dotdash line) when the fitted models are (a) one-sine, (b) song-sine, (c) two-sine with and (d) without damping, respectively

Table 4 Standard error estimates and CI's of the circadian period in a real gene expression dataset

Fitted model	CLS		TS		GLS		MTS	
	95% CI	$\hat{\sigma}$	95% CI	$\hat{\sigma}$	95% CI	$\hat{\sigma}$	95% CI	$\hat{\sigma}$
one-sine	24.15 ± 1.89	35.08	26.50 ± 1.63	29.88	26.23 ± 2.32	30.51	26.55 ± 1.82	29.97
song-sine	23.97 ± 1.45	37.46	26.89 ± 1.73	29.92	27.01 ± 1.89	29.99	27.05 ± 1.93	30.02
two-sine with damping	24.89 ± 2.48	36.05	26.45 ± 1.46	29.24	26.42 ± 2.21	29.74	26.52 ± 1.66	29.32
two-sine without damping	24.04 ± 1.84	38.30	24.24 ± 1.55	34.73	24.25 ± 1.61	34.75	24.20 ± 1.45	34.70

The analysis results indicate that for all the fitted models, the CLS estimates of the circadian periods are approximately 24 h with residual standard errors that are larger than those obtained using TS, MTS and GLS. This is substantiated by the plots of the fitted models showing that CLS produces a slightly poorer fit, as shown in Figure 2. Except for the two-sine without damping model, the other three estimation methods produces circadian period estimates that are larger than 24 h. For choosing the best model and method, Akaike Information Criterion (AIC) is one of the most widely used methods. This is defined as $\mathrm{AIC} = 2k + (rn)\ln\left(\frac{\sum_{i=1}^{r}\sum_{j=1}^{n}\left(y_{ij} - \mathrm{f}(t_{ij};\hat{\boldsymbol{\theta}})\right)^2}{rn}\right)$ where k is the number of parameters in each model fitted by each method. AIC for the four models and methods are shown in Table 5. The CLS method has largest AIC for all models while there are only small differences

between AIC values for TS, GLS and MTS methods. The best fit to the data, clearly supported by the TS, GLS and MTS methods, is the two-sine with damping model.

Table 5 AIC values for each model of the gene expression dataset

Fitted model	CLS	TS	GLS	MTS
one-sine	1,487.92	1,419.35	1,427.88	1,420.49
song-sine	1,516.17	1,420.91	1,421.74	1,422.06
two-sine with damping	1,495.92	1,412.35	1,419.12	1,413.32
two-sine without damping	1,520.17	1,482.87	1,482.92	1,482.34

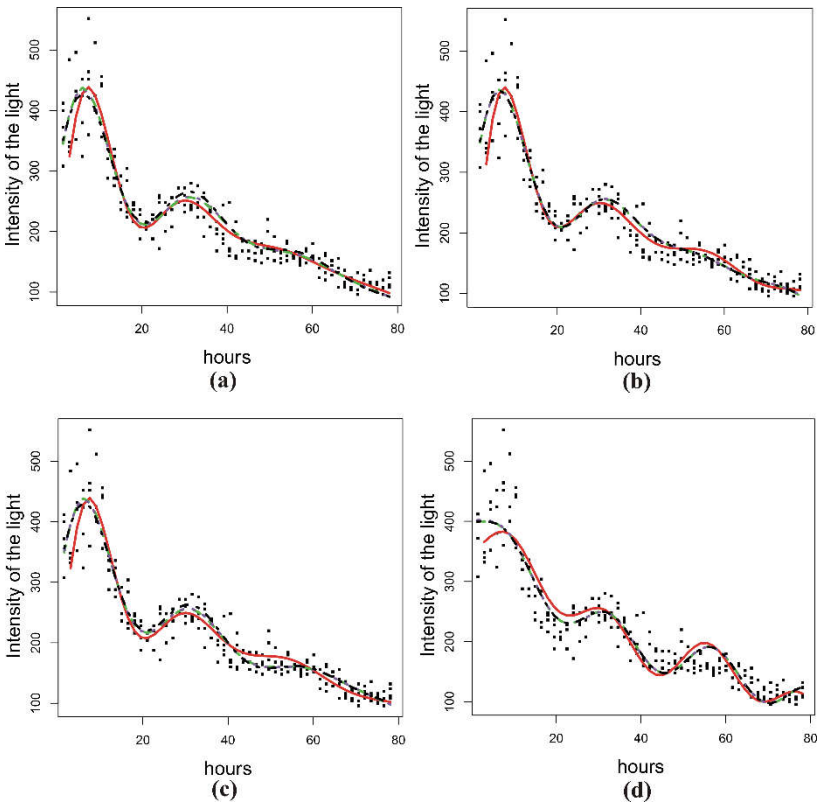

Figure 2 Gene expression observations are fitted by (a) one-sine, (b) song-sine, (c) two-sine with and (d) without damping using CLS (solid line), TS (dashed line), MTS (dotted line) and GLS (dotdash line) procedures

5. Conclusions

In this paper, the modified two-stage method (MTS) is developed to improve coverage probabilities by using pure errors to compute the correlation coefficient in the weight matrix. The modified method is compared to conditional least squares (CLS), two-stage (TS) and generalized least squares (GLS) estimation methods for analyzing circadian rhythm in gene expression data. Simulation results suggest that these methods produce unbiased estimators of the circadian period. The TS method produces poorer confidence intervals than that of CLS. Although GLS is slightly preferred to MTS, in terms of both good variance estimates and confidence intervals, GLS has a

higher failure to converge rate in the iterative fitting process, particularly for the song-sine model. It is not obvious why this is the case and is worth exploring. Failure will lead to unbiased but imprecise results and can also occur in practice. In addition, almost all results of the residual standard errors and Akaike Information Criterion (AIC) show that MTS, TS and GLS models provide a slightly better fit than CLS. Hence, the work here suggests that use of the MTS method can produce reliable estimates and confidence intervals comparable with GLS and, importantly, is more likely to produce a result.

Acknowledgements

We would like to thank the Ministry of Science and Technology (MOST) of Thailand and Kasetsart University, Chalermphrakiat Sakon Nakhon Province Campus (KUCSC) for the financial support.

References

Asikgil B, Erar A. Modified two-stage least squares method. ASMDA 2009: Proceedings of the XIII International Conference on Applied Stochastic Models and Data Analysis; 2009 June 30-July 3; Lithuania. Vilnius; 2009. 124-128.

Bates DM, Watts DG. Nonlinear regression analysis and its applications. New York: John Wiley and Sons; 1988.

Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. Stat Med. 2006; 25: 4279-4292.

Crawley MJ. The R book. West Sussex: John Wiley and Sons; 2013.

Davidian M, Giltinan DM. Nonlinear model for repeated measurement data. London: Chapman & Hall; 1995.

Gallant AR, Goebel JJ. Nonlinear regression with autoregressive errors. J Am Stat Assoc. 1976; 71(356): 961-967.

Izumo M, Johnson CH, Yamazaki S. Circadian gene expression in mammalian fibroblasts revealed by real-time luminescence reporting: Temperature compensation and damping. Proc Natl Acad Sci USA. 2003; 100(26): 16089-16094.

Izumo M, Sato TR, Straume M, Johnson CH. Quantitative analyses of circadian gene expression in mammalian cell cultures. PLOS Comput Biol. 2006; 2(10): 1248-1261.

Klimko LA, Nelson PI. On conditional least squares estimation for stochastic processes. Ann Stat. 1978; 6(3): 629-642.

Kyriacou CP, Hall JC. Circadian rhythm mutations in Drosophila melanogaster affect short-term fluctuations in the male's courtship song. Proc Natl Acad Sci USA. 1980; 77(11): 6729-6733.

Maier B, Wendt S, Vanselow JT, Wallach T, Reischl S, Oehmke S, Schlosser A, Kramer A. A large-scale functional RNAi screen reveals a role for CK2 in the mammalian circadian clock. Gene Dev. 2009; 23(6): 708-718.

Park RE, Mitchell BM. Estimating the autocorrelated error model with trended data. J Econometrics. 1980; 13(2): 185-201.

Pinheiro JC, Bates DM. Mixed-effects in S and S-PLUS. New York: Springer; 2000.

Pukdee W, Polsen O, Baksh MF. Improved methods for the analysis of circadian rhythms in correlated gene expression data. Songklanakarin J Sci Technol. 2018; 40(3), 692-700.

R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2013 [cited 2015 Mar 3]; Available from: http://www.R-project.org/.

Ritz C, Streibig JC. Nonlinear regression with R. New York: Springer; 2008.
Seber GAF, Wild CJ. Nonlinear regression. New York: John Wiley and Sons; 2003.
Yang R, Su Z. Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation. Bioinformatics. 2010; 26(12): 168-174.